Criminal Law Practitioner

Volume 3 | Issue 1 Article 4

2015

Visualizing DNA Proof

Nicholas L. Georgakopoulos
Indiana University

VISUALIZING DNA PROOF

by Nicholas L. Georgakopoulos

Abstract: DNA proof inherently involves the use of probability theory, which is often counterintuitive. Visual depictions of probability theory, however, can clarify the analysis and make it tractable. A DNA hit from a large database is a notoriously difficult probability theory issue, yet the visuals should enable courts and juries to handle it. The *Puckett* facts are an example of a general approach: A search in a large DNA database produces a hit for a cold crime from 1972 San Francisco. Probability theory allows us to process the probabilities that someone else in the database, someone not in the database, or the initial suspect, Baker, may be the perpetrator and obtain the probability of Puckett's guilt. Given the clarity of this analysis, decisions that do not follow it deserve reversal as clearly erroneous.

I. INTRODUCTION

A disease test with 90% accuracy is actually accurate less than 10% when the incidence of the disease is 1%. My guess that the prize is behind the second of three doors, followed by the game host giving me the information that the prize is *not* behind the first door (information that appears pointless) has *half* the chance of success of the alternative, switching my selection to door three. These statements, which are borderline nonsensical, are actually true. They capture two of the several paradoxes of probability theory.[2] Criminal trials on the basis of identifications from large DNA databases are not quite as paradoxical but getting our heads around their probability theory is a monumental task.

So limited seems our ability that I have formed the belief that our difficulty with probabilistic analysis is part of human nature, the result of evolution.[3] No surprise

1 Harold R. Woodard Professor of Law, Indiana University Robert H. McKinney School of Law, Indianapolis, ngeorgak@iu.edu. Special thanks go to Barry Nalebuff who engaged our discussion wholeheartedly. I also wish to thank John Donohue, David Kaye, Erin Murphy, Richard Posner, and Eric Talley for helpful comments and Susan David DeMaine for exceptional librarian and editing assistance.

2 The first is the rare disease or false positive paradox and the second is the three door or Monty Hall Paradox. *See generally* M.H. Rheinfurth and L.W. Howell, *Probability and Statistics in Aerospace Engineering*, NASA, http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19980045313.pdf (April 26, 2015).

3 Perhaps a mutation that facilitated probabilistic analysis appeared in some early hominids, but those went neither to hunt the sabretooth tiger nor to gather fruit in its habitat. Those with the mutation giving a good sense for probability theory, I posit, did not explore new lands, seas, or technologies. They did not write poems and songs about unrequited love. They settled and were selected out of existence by the hunters, the explorers, and the starry-eyed romantics. Perhaps, understanding probability analysis is an evolutionarily unfit trait that

comes from realizing that probability theory developed at about the same time as the calculus because it is about as unnatural for our thinking.[4] The mode of analysis necessary to evaluate DNA evidence from a large database is even more recent, dating from the publication of Bayesian analysis in 1763.[5] The counterintuitive nature of probability theory is especially evident when courts seek to assess the probative value of DNA evidence when the source of that evidence is a large database.[6] DNA databases are enormous and the accuracy of the test presents odds ratios involving numbers well over a million.[7]

Besides the visualizations, the contribution of this analysis is that it proposes the correct analysis when a DNA match arises from the trawl through a large database. The National Research Council has proposed two different adjustments to the random match probability but both have inadequacies, waste information, and do not take advantage of the surrounding environment of the criminal identification.[8]

Part II introduces visualizing with the rare disease test. Part III lays the foundation for visualizing the typical problem presented in *People v. Puckett*,[9] where Puckett was convicted in 2008 for a 1972 rape-murder on the basis of DNA evidence and an investigated suspect, Baker, had not been prosecuted. The generality of the setting is important: The analysis applies in every case of a perpetrator identification through DNA testing of a large database. Part IV visualizes the three possible scenarios that the early suspect was the perpetrator, that the perpetrator was not in the database, and that the perpetrator was in the database. Part V produces the corresponding probability tree, and Part VI does the number crunching to calculate the probability of Puckett's guilt, which turns out to be almost 99%. The conclusion circles back to the treatment of evidence that would allow the courts to perform the probability theory analysis.

II. The Rare Disease Paradox

Suppose a disease infects one percent (1%) of the population, and a relatively accurate test exists for this disease, one that has 90% accuracy. Importantly, accuracy

we cannot have.

4 *See generally* R.R., *The Discovery of Calculus,* Science Reviews 2000 Ltd. (1919), http://www.jstor.org/stable/43427110 (April 26, 2015) (Isaak Newton and Gottfried Leibnitz discovered the calculus simultaneously around 1666 to 1684.).

5 *See generally* Roger North, *The Mathematical Gazette*: *The Mathematical Career of Pierre de Fermat by Michael Sean Mahoney*, Mathematical Association (1974), http://www.jstor.org/stable/3616110 (April 26, 2015) (demonstrating that modern rigorous probability theory dates from correspondence between Pierre de Fermat and Blaise Pascal in 1654); Joseph Berkson, *Bayes' Theorem*, The Annals of Mathematical Statistics (1930), http://www.jstor.org/stable/2957673 (April 26, 2015) (stating that the Bayesian analysis applied to this issue dates from 1763).

6 David H. Kaye, *Rounding up the Usual Suspects: A Legal and Logical Analysis of DNA Trawling Cases*, 87 N. Car. L. Rev. 425 (2009) (offering an eloquent overview of the courts' attempts to deal with large database DNA evidence).

7 *See* Ian Ayres & Barry Nalebuff, *The Rule of Probabilities: A Practical Approach for Applying Bayes' Rule to the Analysis of DNA Evidence*, 67 Stanf. L. Rev. 1447 (2015) (noting the complexity of DNA analyses).

8 The National Research Council has suggested two adjustments. In its first report, it recommended that database searches only use a few of the places (loci) where human DNA has the differences that are used for identification and after the search reveals a suspect, that suspect's identification proceed on the basis of the remaining of the 13 loci that the database holds. For example, the database search uses data of 8 of the loci from the sample at the crime scene to identify a suspect; then, the remaining 5 loci confirm the suspect's identity. The second report suggested that the odds ratio of the test's error be multiplied by the size of the database. For example, if the test errs once in a billion, and the database has one million members the error rate becomes one million in one billion or one in a thousand. *See* Kaye, *supra* note 5, at 436-43; Comm. on DNA Tech. in Forensic Sci., Nat'l Research Council, *DNA Technology in Forensic Science*, 124 (1992) ("NRC I"); Comm. on DNA Forensic Sci.: An Update, Nat'l Research Council, *The Evaluation of Forensic DNA Evidence* 134 (1996) ("NRC II").

9 *People v. Puckett*, No. SCN 201396 (Cal. Super. Ct. Feb. 4, 2008). *See generally* Kaye, *supra* note 5; Ayres & Nalebuff, *supra* note 6 (citing *People v. Puckett*).

"A DISEASE TEST WITH 90% ACCURACY IS ACTUALLY ACCURATE LESS THAN 10% WHEN THE INCIDENCE OF THE DISEASE IS 1%."

means that the test both identifies infected individuals with 90% probability (what some disciplines call sensitivity, true positive rate, or recall rate) and identifies healthy individuals with 90% probability (this aspect of accuracy some disciplines call specificity, or true negative rate), or conversely, fails to identify them as healthy with 10% probability. The paradox appears when we posit that an entirely random individual receives a positive result, a result that flags this person as infected. The usual lay intuition is that this person's infection probability is near 90%, but the actual probability of infection is under 10%. What drives this discrepancy between our intuition and the accurate calculation is that our intuition does not account for false negatives: the frequency with which the test flags healthy subjects as infected. The accurate calculation requires us to realize that because the uninfected population is so large, the proportionately few false positives they will receive are actually many in comparison to the few true positives of the tiny infected fraction of the population.

A visual representation of the paradox illustrates the accurate approach. Consider Figure 1, a grid of one thousand dots, ten of which, 1%, are black and the rest are white. This represents the reality of a population with 1% infected individuals.

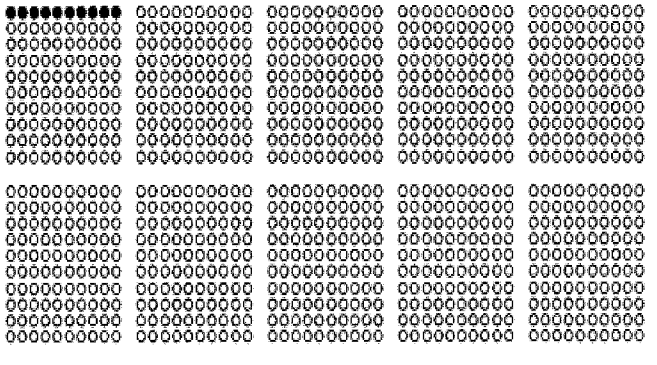


Figure 1: A grid of 1,000 dots, 1% of which are black, that corresponds to the paradox of the rare disease test. This is the true state of the world to which an imperfectly accurate test is applied.

When we apply the test for the rare disease to this population, the result contains errors. The errors take two forms, false negatives and false positives. A false negative occurs when the test of an infected individual (one of the black dots in Figure 1) flags that person as uninfected, as a white dot. A false positive presents an uninfected individual as infected. Figure 2 has randomly flipped the color of one dot in each row of ten, producing a 10% error in the observations of the true state of the dots from Figure 1.

Once we visualize the false positives, their frequency becomes apparent. An individual receives a positive test. How probable is it that this positive result is one of the infected dots versus the false positives?

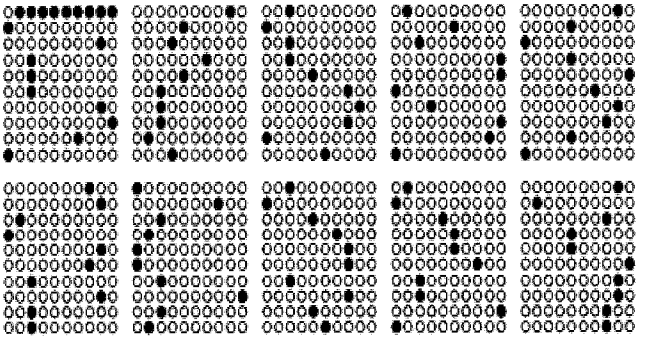
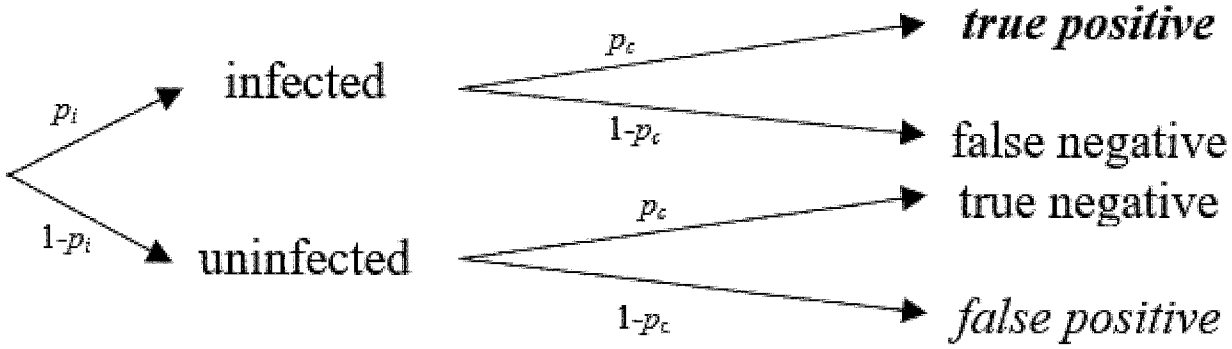


Figure 2: The 10% error rate of the test reverses one dot in each row of ten.

Only nine true positives exist in a sea that includes ninety-nine false positives. Given a black dot, the probability that it is true is nine in one hundred eight (the total number of black dots), under ten percent (actually 8 ⅓ %), despite the test's ninety percent accuracy.

Figure 3: The probability tree that corresponds to the paradox of the rare disease test. The probability of a positive being true has as its denominator the sum of the probability weights that correspond to all positives, the italicized endpoints. Its numerator is the probability weight that corresponds to the true positive, the bolded endpoint.



To confirm the accuracy of this analysis, let us also visualize it as a probability tree, as in Figure 3. To calculate the probability of infection given a positive signal, we must account for all possibilities of observing a positive, which are two, a true positive and a false positive, the italicized endpoints of the probability tree. The denominator must hold the sum of the probability weights that correspond to all positives. In this case, the true positive occurs when a subject is infected ($\boldsymbol{p}_i$ = 1%) and the test is correct ($\boldsymbol{p}_c$ = 90%), for probability weight of .01 × .90 = .009. The false positive occurs when a subject is not infected ($1 - \boldsymbol{p}_i$ = 99%) and the test is false ($1 - \boldsymbol{p}_c$ = 10%), for a probability weight of .99 × .10 = .099. The sum of those two, .108, is the denominator. The numerator is the first of the two, the probability weight that corresponds to a true positive, the endpoint of the probability tree that is in bold (as well as in italics). That is .01 × .90. The result is the same 8⅓% calculated in the graphical approach. Table 1 presents this calculation.

Case:	Calculation:
True Positive:	$p_i p_c$ = .009
False Positive:	$(1-p_i)(1-p_c)$ = .099
Numerator:	.009
Denominator:	.009 + .099 = .108
Probability:	.009/.108 = .083

Table 1: The probability weights of each case of a positive in the rare disease test leading to the calculation of the probability that a positive is a true positive.

The DNA test in *Puckett* is more complex, but the principle is the same. We receive a signal, that is, we see a black dot or a positive DNA test. We need, first, to determine the universe of black dots, true and false. Second, we must calculate the probability that this signal corresponds to a true black dot, i.e., a correctly convicting DNA test. But, just as detectives must start at the crime scene, we must start with the San Francisco of 1972.

III. Visualizing 1972 San Francisco

Over 40 years ago, twenty-two year old Diana Sylvester was found dead in her apartment in San Francisco.[10] She had been raped and murdered a few days before Christmas 1972.

In 2003, California police check a preserved DNA sample against the California database containing felons' DNA. John Puckett is a match. What is the probability that he is guilty? The setting presents a similar paradox to the rare disease case in the sense that the accuracy of the test is very large but applying it to a database of that size would produce a false positive with significant probability.[11]

The first layer of complexity is that the match comes

10 *See* Ayres & Nalebuff, *supra* note 6, at 1467-68; Michael Bobelian, *DNA's Dirty Little Secret*, Washington Monthly (March/April 2010), http://www.washingtonmonthly.com/features/2010/1003.bobelian.html.

11 The error rate of the test according to the prosecution's expert was one in 1,100,000, meaning that one person in 1,100,000 individuals who were not the sources of the DNA would have the same DNA sequence ("random match probability"). Applied to the database that had 338,711 elements produces a random false positive with about 26.5% probability. *See infra* note 12. *See generally* Erin Murphy, *The Art in the Science of DNA: A Layperson's Guide to the Subjectivity Inherent in Forensic DNA Typing*, 58 Emory L.J. 489 (2008) (discussing the mechanics of DNA identification and its excessive purported accuracy including excellent graphics).

from the DNA database, but the suspect must come from the people who were in San Francisco at the time of the crime in 1972 and were of rape-committing age. For simplicity, I will call this the [population of] 1972 San Francisco. Most entries in the database are not from 1972 San Francisco.

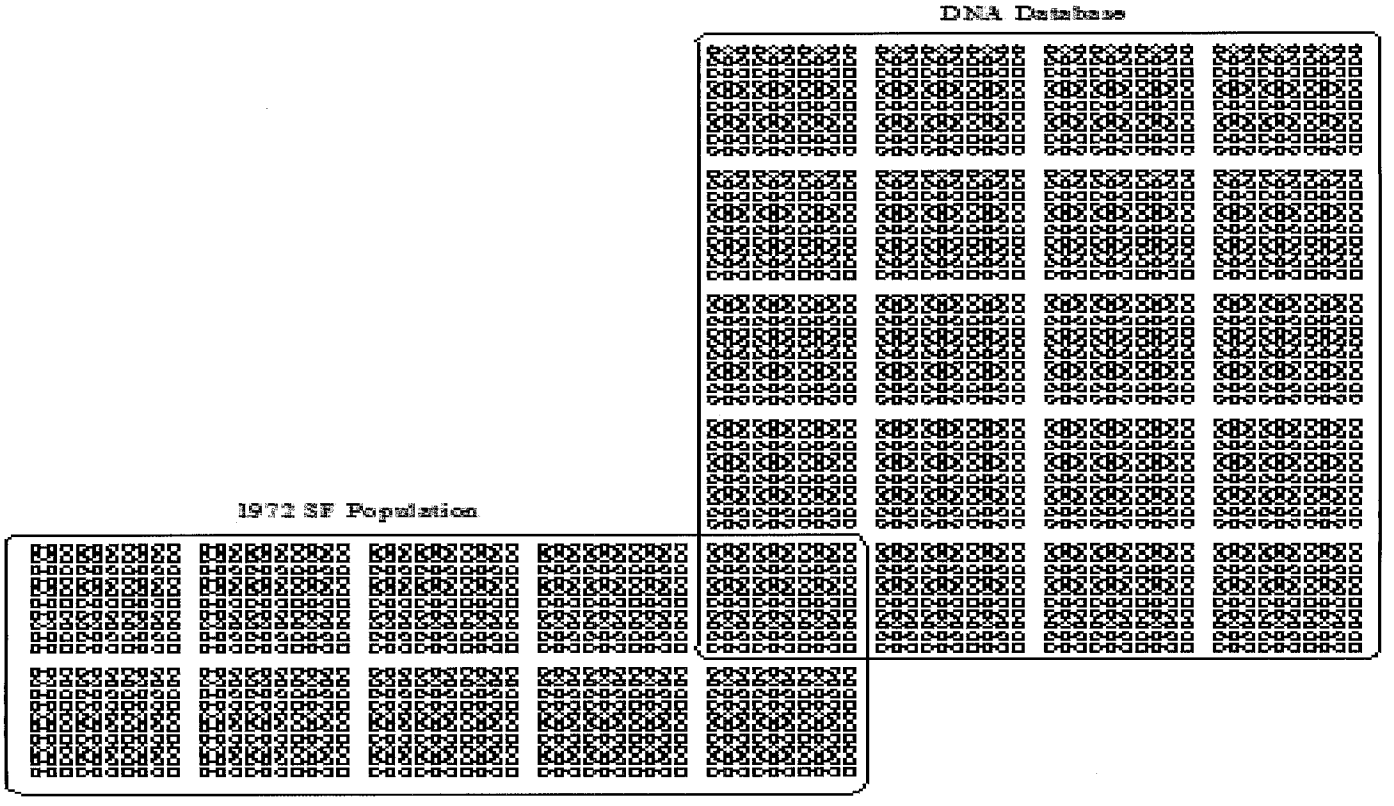
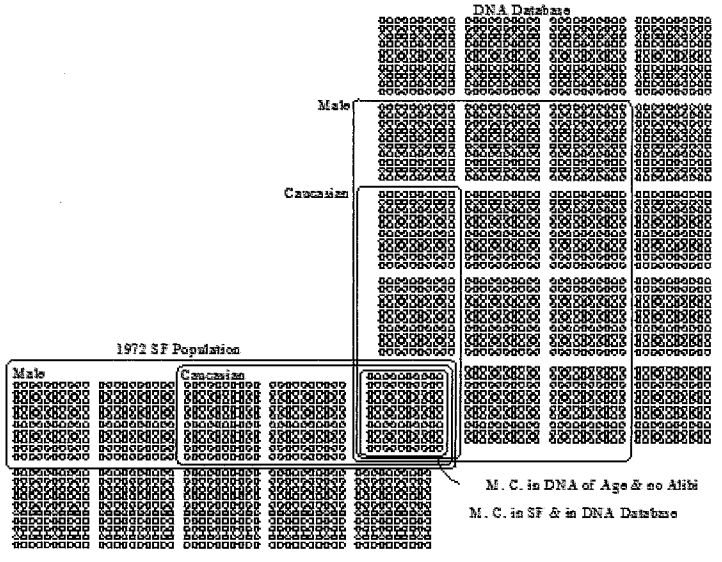


Figure 4: An illustration of the population of 1972 San Francisco (stylized to one thousand) and its intersection with the relevant population of the DNA database (stylized to two thousand). The intersection holds one hundred elements, ten percent of the population of San Francisco and five percent of the database.

For the visualization, let us temporarily reduce the size of the population of 1972 San Francisco to one thousand and the size of the database to two thousand. Suppose that their intersection, their relevant common elements, holds a hundred members. Figure 4 illustrates this overlap.

Assessing the probative power of a DNA test depends on the size of the intersection, the population both in the DNA database and in San Francisco at the time of the crime. One approach is to estimate the fraction of the database that is male, then the fraction that is Caucasian (because a witness saw a Caucasian man in the victim's apartment), and, finally, the fraction that was of rape-committing age in 1972 and has no alibi.[12] Figure 5 illustrates this approach by circling successively smaller fractions of the database. The figure also illustrates the alternative approach to estimating the intersection: by taking successively smaller fractions of the San Francisco population.[13] These correspond to its male fraction, its Caucasian fraction, and, finally, its fraction on the



12 This is, simplified, the approach that Ayres and Nalebuff use. *See generally* Ayres & Nalebuff, *supra* note 6.

13 This is analogous to the simpler estimation based on the Bay area population that Kaye uses. *See generally* Ayres & Nalebuff, *supra* note 6.

IV. The Three Alternatives

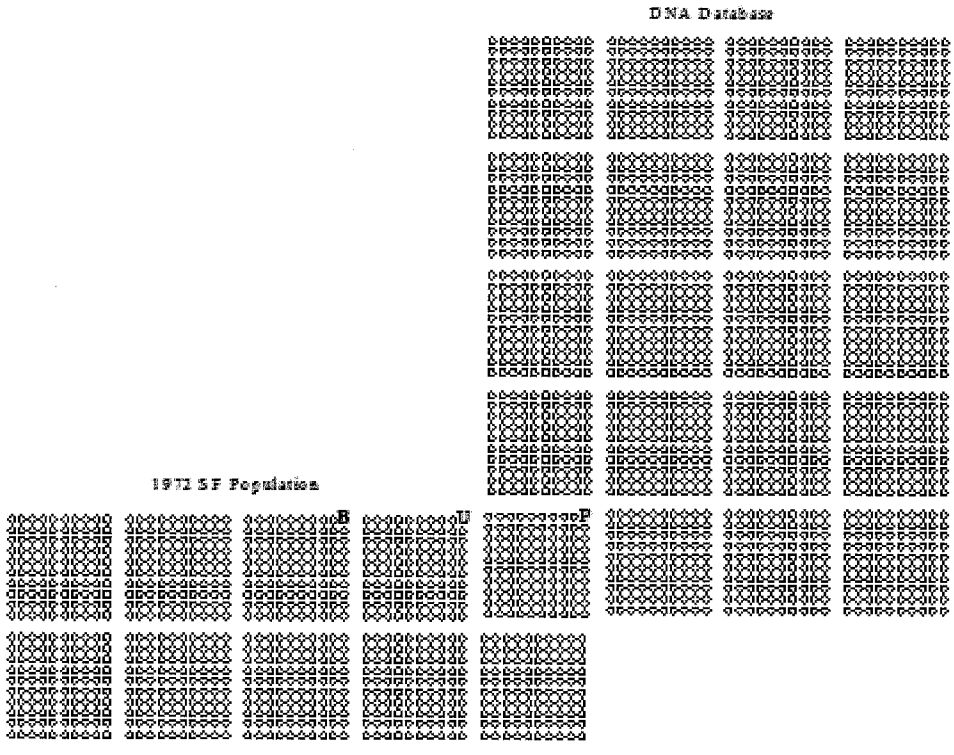


Figure 5: Different approaches to estimating the overlap of the population and the database.

The method of approaching the estimate of the intersection changes the inquiry in intuitive ways. For example, one who starts from the database needs to ask what fraction of the database was of age in 1972. Also, the fraction male and Caucasian has, then, as its denominator the database population. By contrast, one who approaches the estimate from the 1972 San Francisco population has already excluded implicitly individuals who are too young for the crime in 1972 and those with an alibi of being elsewhere. Also, the fractions of males and Caucasians that matter are those of San Francisco, i.e., their denominator is the 1972 population of San Francisco.[14]

In sum, the first issue is estimating the population at the intersection of the population and the database. The next hurdle is to identify the possible alternative perpetrators.

Figure 6: Baker (B), unknown (U), and Puckett (P) as possible positions relative to the intersection of the San Francisco population and the DNA database.

The alternative perpetrators are three: The perpetrator may be Baker, the lead suspect at the time, the perpetrator may be an unknown not in the database, or someone in the database (who most likely is Puckett unless the perpetrator received an unlikely false negative and Puckett a false positive). Baker died in 1978 without leaving a DNA sample.[15] If Baker was the perpetrator, Puckett received a false positive. Similarly, Puckett received a false positive if the perpetrator was an unknown who is not in the database. Finally, the perpetrator may be in the database, in which case we are most likely observing a correct identification of Puckett as the perpetrator but the possibility exists that Puckett is a false positive that arises after the true perpetrator received a false negative.

Figure 6 illustrates these three alternatives by identifying three points with B, U, and P. The location of the three points is significant. The first two, B (Baker), and U (the unknown) lie in that part of the population of San Francisco that corresponds to the subset that is male and Caucasian but outside the subset that overlaps with the DNA database. Puckett's P, on the other hand, is in the intersection.

14 For example, based on census data one could estimate the 1972 San Francisco population at 720 thousand, its Caucasian fraction at 60%, and take the fraction with which Caucasians end in the felons' DNA database at about 2%, to produce an estimate of the intersection of about 720,000 × .6 × .02 = 8,640. This is quite close to the estimate formed by the method of Ayres and Nalebuff of about 8,790, *see infra*, text following note 27. Kaye approximates this intersection by using the 2003 population of the entire Bay area to about 2 million. *See* Kaye, *supra* note 5 at 491. If he were to reduce that to the proportion Caucasian, say 50%, and in the database, 2%, that would yield an intersection of about 50,000, still far larger, but likely near the maximum that the defense could plausibly argue to be reasonable.

15 Ayres & Nalebuff, *supra* note 6 at 1487.

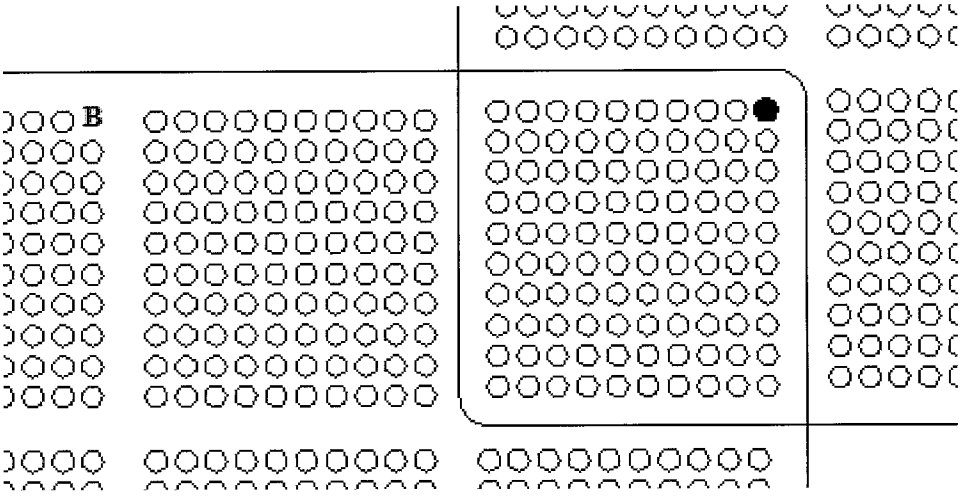


Figure 7: The first possibility is that Baker is the perpetrator and we observe a false positive. The false positive arises in the intersection, the shared elements between the San Francisco population and the DNA database. Baker, identified with a B, is not in that subset but is part of the male and Caucasian subset of the San Francisco population.

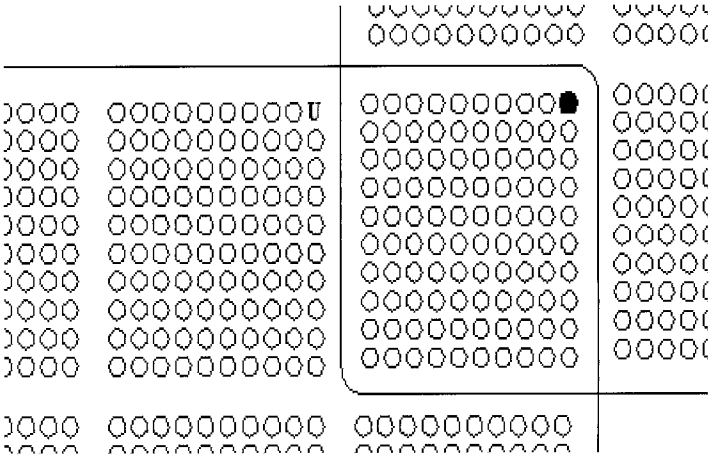


Figure 8: The second possibility is that an unknown individual U, not in the database, committed the crime and we are observing a false positive.

The first possible world is the one where Baker was the perpetrator and we observe a false positive from the DNA database. Figure 7 illustrates this world. One of the hundred points at the intersection of the San Francisco population and the DNA database is black and corresponds to the false positive.[16] We see with a B the location of Baker. While this visualization shows one black dot in a hundred, the corresponding exact calculation comes in Part V, with the probability tree, figure 10.

The second possible world is where the perpetrator was an unknown person who is not in the database. Figure 8, not very different from the previous picture, illustrates this alternative. The U denotes the unknown person who committed the crime. This unknown person is male and Caucasian, but is not in the database. One of the points in the intersection of the population and the database appears black as a false positive.

The third and last alternative is that the perpetrator is in the database. One might think that Puckett corresponds to a single black dot but that is wrong because Puckett's guilt is a virtually certain phenomenon in this third alternative. For the purpose of the illustration, Puckett's point is the entire intersection: If Puckett is guilty we almost always see a true positive with the unlikely exception of a false negative that exonerated the perpetrator, followed by an also unlikely false positive fingering Puckett. To visualize the corresponding almost 100% probability of having identified Puckett in juxtaposition with our prior rare false positives, superimpose the 100% reality on the intersection in those same graphs to see the intersection as mostly black dots. Thereafter we can see the possibility that the true perpetrator experiences the rare false negative by leaving white a dot (or a fraction of one) corresponding to the probability of a false negative.[17]

16 Figure 7 shows one of the hundred dots at the intersection as black. This does not correspond to a test with 99% accuracy but rather to one with accuracy of 99.9899502%, because $99.9899502^{100} = .99$. DNA tests generally have much greater accuracy, with error rates measured as one in billions. In Puckett's case, the naïve position that the positive was merely the result of applying it to the entire database of 338,711 samples gives the impression that the probability of a false positive was the accuracy of the test, 1,099,999/1,100,000, raised to that power, which gives a probability of producing that number of correct negatives was slightly under 73.5% and, therefore, the probability of false positives slightly over 26.5%.

17 In the setting of this visualization letting an entire dot be white strongly overstates the prob-

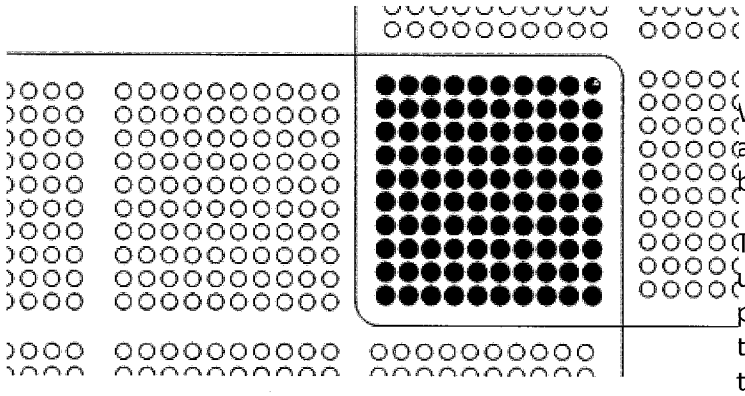


Figure 9: Puckett's guilt occupies most of the probability space, which takes the shape of the intersection of the San Francisco population and the database to be comparable to the alternatives. The only caveat is a false negative, but we can visualize it as a partially white dot, one tenth white for a test with 99.9% accuracy. The fractional filling of about a hundredth of the white space corresponds to a false positive after a false negative.

One possibility still remains. In the case that the true perpetrator is in the database and surprisingly receives a false negative, then the remaining members of the intersection of the population and the database might not all receive true negatives and a false positive may still arise. In terms of the visualization, a very small fraction of the (likely partially) empty dot that signified Puckett's false negative is black. That, however, must be accounted in the probability space of false positives. In other words, the fraction of the dot that can arise as a false positive after the perpetrator receives a false negative should be added to the probability weight that the first two figures produce and which corresponds to false positives.

Unlike the disease setting, where many black dots were associated with false positives, here the odds favor the true positives. The visual, stylized representation of the *Puckett* setting gives us 99.9 true positives versus two and a very small fraction of false positives, while considering (i) the three scenarios as equally likely and (ii) the test to have accuracy that produces ninety nine true negatives in a hundred. The DNA tests are a lot more accurate, the estimated probabilities of the three scenarios are unequal, and the analysis needs to remain sensitive to changes of the estimates of the various probabilities. We need to construct the corresponding probability tree.

V. Probability Tree

While the rare disease test produced a simple probability tree, the trial setting produces a complex one because of the several uncertainties.

The initial branching corresponds to the most general uncertainty, whether a different suspect was the true perpetrator, who in *Puckett* was Baker. This forms the initial branching between the probability p_B that this other (Baker) was the perpetrator and $1-p_B$ that he was not.[18] If Baker was not the perpetrator, the next uncertainty is whether the perpetrator was in the intersection of the DNA database and the population. The corresponding branching is that the perpetrator was in the database with probability p_d and was not in it with probability $1-p_d$.

From (1) the node corresponding to another (Baker) being the perpetrator and from (2) the node corresponding to the perpetrator not being in the database, the subsequent branching is identical because in both cases any positives are false positives and the intersection of the database and the population holds the same number of members, N. The branching is triple, with the first case being that all members receive correct negative tests.[19] The test correctly rejects a DNA match with probability r.[20] Because all members must receive a true

ability of a false negative. In the prior two figures the number of black dots was one, implying that the test's accuracy is 99.9899502% (*See supra* note 12). To be consistent, about one hundredth of a dot should be white here.

18 The same analysis applies if more than one alternative suspect exist. The probability assigned to Baker in this example would need to be adjusted to include the cumulative probability of all other suspects. If the two alternative suspects, for example, were Able and Charlie, with Able having a 20% probability of being the perpetrator and Charlie a 5% probability, the appropriate value of p_B would be .25.

19 A simpler analysis merely bifurcates here between everyone receiving true negatives with probability r^N or not, $1-r^N$. This produces the probability tree for one or more positives, however. At sample sizes like this one, where much less than one false positive is expected on average, this calculation is not very different, as table 2 and note 28 show and as Part VII explains. *See infra* note 34.

20 This is the rate of accuracy of the test, also known as the true negative rate or specificity of the

negative, the operation is multiplicative. Say ***r*** were .90, for simplicity. Ninety percent of the time, then, the first test would be negative. The second would also be negative ninety percent of ninety percent of the time or .90 squared, and the third also ninety of ninety of ninety, or .90 cubed. Accordingly, the probability of all ***N*** members receiving correct negative tests is the accuracy of the test ***r*** raised to the power of the number of members of the intersection, for probability r^N.

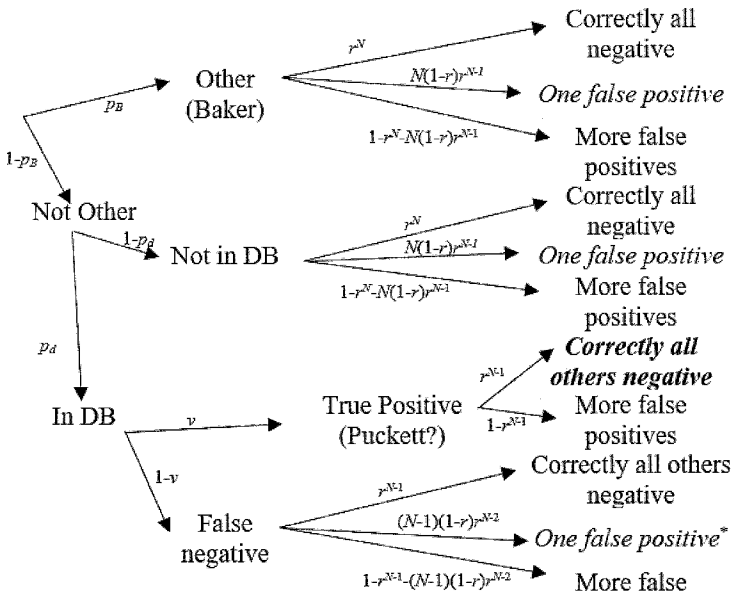


Figure 10: The probability tree for exactly one positive. The italicized endpoints correspond to observations of exactly one positive. The sum of their probability weights forms the denominator of the probability of Puckett's guilt with the numerator being the probability weight that corresponds to the true positive, the bolded endpoint. Ayres and Nalebuff treat the starred false positive in a footnote, justifiably considering it trivial, but the graphical analysis lets it remain in the foreground.

The second branch of the triple branching corresponds to exactly one member of the intersection receiving a false positive. Again, the calculation is easier to see in a simplified setting. Consider three coin-tosses of a biased coin that produces heads with 90% probability. A single tails appears in three sequences of results: tails-heads-heads; heads-tails-heads; and heads-heads-tails. Each sequence has one 10% event (a tails) and two 90% events (two heads) which corresponds to each path having probability $(1-.9).9^2$; keep in mind that if the number of tosses were ***N***, then the number of 90% events would be ***N***-1. The probability that any of the three paths materializes is $3(1-.9).9^{3-1}$. Generalize by replacing 3 with the number of the uncertain events N and .9 with the probability that the DNA test rejects an innocent match, ***r***, to obtain $\boldsymbol{N}(1\boldsymbol{r})\boldsymbol{r}^{N1}$. That is the probability of exactly one false positive. This calculation is also given by the probability density function of the binomial distribution for ***N*** trials with probability of success in each trial 1-***r***.[21] The corresponding intuition has two components. First, one of the ***N*** members must receive a false positive. Since each receives a false positive with probability 1-***r***, this is ***N***(1-***r***). Second, the remaining ***N***-1 members of the intersection must all receive true negatives, which is $\boldsymbol{r}^{N-1}$. The resulting probability that exactly one false positive appears is $\boldsymbol{N}(1\boldsymbol{r})\boldsymbol{r}^{N1}$. The endpoint of this branch appears in italics to signify that it corresponds to the observation of exactly one positive. The sum of the probability weights of all such endpoints forms the denominator of the probability of Puckett's guilt.

The third branch of the triple branching contains the remaining probability weight, one minus the probability of the first two branches. This corresponds to more than one positives appearing and is $1-\boldsymbol{r}^{N}-\boldsymbol{N}(1\boldsymbol{r})\boldsymbol{r}^{N1}$.

From the remaining node that corresponds to the perpetrator being in the database, the first uncertainty is the obvious one, whether the perpetrator will receive the true positive test. Despite that intuition suggests that the probability of a true positive is the same as that of a true negative, ***r***, because different uncertainties may arise, call the probability of a true positive ***v*** (what some disciplines call the true positive rate or sensitivity of the test).[22] Thus, the initial branching will be that

test. Both Kaye and Ayres and Nalebuff focus on the error rate, 1-***r*** in my terms, which would produce slightly different equations, but consistent after one makes the appropriate substitutions.

21 The mathematical knowledge repository www.wolframalpha.com gives this result, for example, if one enters "PDF[BinomialDistribution[n, 1-r], 1]" asking for the value of the probability density function for obtaining one positive from a binomial distribution with n trials with probability of success 1-r.

22 Whereas we have a probabilistic sense of false positives, we do not have a theory of false negatives that is based on the probability theory of DNA analysis because the test describes the DNA, so if both the sample at the crime scene and the sample from the perpetrator come from the same individual, the perpetrator, then the test result will necessarily be a match. Error can arise from sources outside the theory of DNA matching, such as sample contamination through laboratory error. *See* Comm. on DNA Forensic Sci.: An Update, Nat'l

the perpetrator, being in the database, will receive a true positive with probability v, and will receive a false negative the rest of the time, $1v$. Additional positives may appear, however, and the probability tree needs to exclude them.[23] This happens by having a branching after the true positive for either all remaining N-1 members of the intersection receiving true negatives, with probability r^{N-1}, or not, with probability $1-r^{N-1}$. The first of these endpoints corresponds to observing exactly one positive and, therefore, is in italics. Because this is the true positive, this endpoint is also in bold and its probability weight will be the numerator of the fraction that gives Puckett's guilt.

After a false negative, again a triple branching appears.[24] First, the remaining members of the database, $N1$, will all produce true negatives with probability r^{N-1}. Second, exactly one false positive will appear, in a way analogous to Baker being the perpetrator but here the intersection is smaller by one member. The single false positive appears with probability $(N-1)(r-1)r^{N-2}$. This is the case where exactly one positive appears and, therefore, is in italics in the figure. The rest of the time, $1-r^{N-1}-(N-1)(r-1)r^{N-2}$, two or more false positives may appear.

Figure 10 displays the probability tree that results from this analysis. The initial node is at the top left and eleven endpoints appear on the right side. The four italicized endpoints correspond to observing one positive and three of those correspond to observing a false positive. The italicized endpoint that is also bold corresponds to observing exactly one positive and that positive being true. The probability of Puckett's guilt has as its denominator the sum of the probability weights that correspond to all four italicized endpoints. The numerator is the true positive, the endpoint that is also bold.

VI. Number Crunching

The return from imagery to arithmetic requires us to put numbers on various parameters. The accuracy (true negative rate or specificity) of the DNA test is r = 1,099,999/1,100,000 = .99999909,[25] the size of the DNA database is D = 338,711,[26] the probability that the suspect is in the database is p_d = .6.[27] The fraction of the database that is male is l = .86 and the fraction Caucasian is c = .284.[28] The fraction of age is g = .425.[29] Taking further fractions of the database, the fraction not incarcerated is n = .67,[30] the fraction without an alibi is o = .5,[31] and the fraction of the database that is not duplicated is s = .75.[32] The prior probability of Baker's guilt is p_B = .3.[33] The true positive rate v is assumed equivalent to the true negative rate, r.

Research Council, *The Evaluation of Forensic DNA Evidence* 134 (1996) ("NRC II") (explaining that it cannot propose such a probability of error):

> There has been much publicity about ... errors made by Cellmark in 1988 and 1989, the first years of its operation. Two matching errors were made in comparing 125 test samples, for an error rate of 1.6% in that batch. The causes of the two errors were discovered, and sample-handling procedures were modified to prevent their recurrence. There have been no errors in 450 additional tests through 1994. Clearly, an estimate of 0.35% (2/575) is inappropriate[ly high] as a measure of the chance of error at Cellmark today.
>
> Rather, the implied error rate should be much smaller, especially assuming the recommended safeguards that include repeat testing by different laboratories.

23 The simpler analysis for one or more positives would not need to exclude additional positives and would not have this branching.

24 Again, the simpler analysis for one or more positives would have a bifurcation here, between all N-1 remaining members of the intersection receiving true negatives with probability r^{N-1}, and not, with probability $1-r^{N-1}$.

25 *See* Ayres & Nalebuff, *supra* note 6 at 1476. Note that the symbol r here is the accuracy of the test, whereas Ayres and Nalebuff use r to symbolize the error rate, what in my terms is $1-r$.

26 *Id.* at 1470.

27 *Id.* at 1479. Up to here the symbols coincide with those of Ayres and Nalebuff but for this they use p rather than p_d. They do not assign symbols to the subsequent variables.

28 *Id.* at 1477.

29 *Id.*

30 *Id.*

31 Ayres & Nalebuff, *supra* note 6 at 1478.

32 *Id.*

33 *Id.* at 1488.

Perpetrator case	Calculation
Other (Baker):	$p_B N(1-r)r^{N-1} = .002\ldots$
Not in the DB:	$(1-p_B)(1-p_d)N(1-r)r^{N-1} = .002\ldots$
In DB, true positive:	$(1-p_B)\,p_d\,v\,r^{N-1} = .416\ldots$
In DB, false positive:	$(1-p_B)p_d(1-v)(N-1)(1-r)r^{N-2} = .000\ldots$
Numerator:	.416...
Denominator:	.002...+.002...+.416...+.000...=.4212...
Probability:	.416.../.4212...=.98909...

Table 2: The probability weights of each case of a positive and the resulting calculation of the probability of Puckett's guilt.[34]

Applying the successively smaller fractions to the DNA database gives the size of the intersection after all the reductions as $\boldsymbol{N = D \times l \times c \times g \times n \times o \times s}$ = 338,711 × .86 × .284 × .425 × .67 × .5 × .75 = 8,789.72. In the context of the illustrations, this is the size of the overlapping population, the intersection of the population of 1972 San Francisco and the population of the DNA database (instead of 100 that figures 7-9 show).

The remaining calculation depends on whether the setting is one where exactly one positive is observed, as in the probability tree of figure 10, or the simpler analysis of one or more positives per notes 9 to 11. Table 2 shows that calculation (note 21 shows the corresponding entries for the simpler analysis). Each row corresponds to one of the ways of observing a single positive, and shows the formula for its probability weight. The last three rows produce the numerical results of the calculations, the probability of Puckett's guilt, which is 98.9% under these assumptions.[35]

VII. Conclusion

That probability theory is difficult and counterintuitive is not news. Rather, the point is that the graphical approach helps make this counterintuitive and very complex analysis comprehensible and the calculations tractable.

The graphical exposition clarifies the analysis. Some argue that juries should evaluate the probabilistic analysis despite its complexity. Hopefully, courts can help juries to handle this complexity. At the very least, however, if juries are to evaluate probability theories, jurors must see the corresponding probability tree and should receive a spreadsheet in which they can see the effect of changing estimates about the inputs into the calculation.

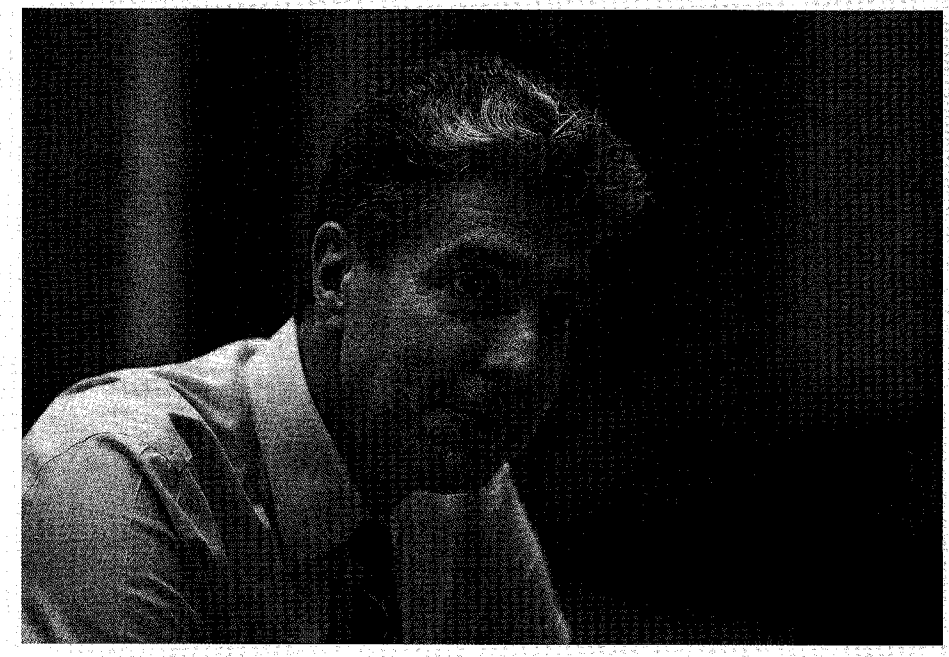
The key point, however, is that the model for analyzing the *Puckett* setting captures the way that cold-hit DNA identifications will tend to arise. In many cases, some initial suspect may keep some probability of still being the perpetrator, as did Baker. Even if such a suspect does not exist, the model still works by putting the corresponding probability ($\boldsymbol{p_B}$) at zero. This is the appropriate analysis rather than the adjustment of the random match probability that the second report of the national research council proposed in 1996. The development of general approach to evaluating DNA evidence means that decisions, like *Puckett*, that ignore this analysis without having truly different facts should be reversible under the clearly erroneous standard.

34 The entries of the simpler probability tree corresponding to one or more positives (per notes 15 to 18) would be as follows: Baker: $p_B(1-r^N)$; Not in DB: $(1-p_B)(1-p_d)(1-r^N)$; In DB true positive: $(1-p_B)\,p_d\,v$; In DB false positive: $(1-p_B)p_d(1-v)(1-r^{N-1})$; numerator: 0.419...; denominator: 0.4246...; probability: .98913... The intuition behind the difference of the two analyses appears if we let $\boldsymbol{N}$ go to infinity. Then, the one or more analysis converges to the probability of the perpetrator not being Baker, being in the database, and receiving a true positive, $(1-\boldsymbol{p_B})\,\boldsymbol{p_d v}$, as many positives appear and one is likely to be the perpetrator. By contrast, the probability of guilt under the exactly one analysis approaches zero, as more positives become exceedingly likely and seeing only one becomes unlikely regardless of guilt.

Spreadsheets of this model are available; Excel: http://tinyurl.com/n4nxdhu; Google docs: http://tinyurl.com/mwr5nna.

35 *See* Ayres & Nalebuff, 67 Stanf. L. Rev. at 1488 (showing the exact same calculation.)

About the AUTHOR



Nicholas L. Georgakopoulos is the Harold R. Woodard Professor of Law at the Robert H. McKinney School of Law of Indiana University. Professor Georgakopoulos specializes in economic analysis of law, mostly business law, with an emphasis to applications of probability theory. His more than forty articles are widely cited by the courts, including the Supreme Court and by the SEC to the Supreme Court. Notable books of his are Principles and Methods of Law and Economics: Basic Tools for Normative Reasoning (Cambridge Univ. Press, 2005) and The Logic of Securities Law (forthcoming, Cambridge U.P. 2016).